# Nan Tang

📱 *+974 66700540*
☎ *+974 44542850*
✉ *ntang@hbku.equ.qa*
🖱 *da.qcri.org/ntang/*
*DOB: 28 September 1980*

*Senior Scientist*
*Qatar Computing Research Institute (QCRI), Doha, Qatar*

## Research Interests

Data preparation theory and systems; deep learning for data preparation; data preparation for machine learning; data visualization for data preparation

## Education

2004/07–2007/12 **Ph.D.**, Systems Engineering & Engineering Management
**`The Chinese University of Hong Kong,`** Hong Kong
• Thesis: Efficient XPath Query Processing in Native XML Databases
• Co-Supervisors: Jeffrey Xu Yu and Kam-Fai Wong

2001/09–2004/01 **M.Sc.**, Computer Science
**`Northeastern University,`** China
• Thesis: Parallel XML Databases
• Supervisor: Guoren Wang

1997/09–2001/01 **B.S.**, Computer Science, **`Northeastern University,`** China

## Professional Experience

2015/04–now *Senior Scientist*, **`Qatar Center for Artificial Intelligence, QCRI`**, Qatar

2017/07–08 *Visiting Scientist*, **`MIT`**, US. Worked on the Data Civilizer project, with Michael Stonebraker, Samuel Madden, and Armando Solar-Lezama

2011/12–2015/03 *Scientist*, **`Data Analytics, QCRI`**, Qatar

2010/02-2012/01 *Research Fellow*, **`University of Edinburgh`**, UK. Worked on data cleaning and graph algorithms, with Wenfei Fan

2008/02-2010/01 *Scientific Staff Member*, **`CWI`** (the national research institute for mathematics and computer science), the Netherlands. Worked on column-store database MonetDB and distributed XQuery processing, with Peter Boncz

2007/03–08 *Visiting Scholar*, **`University of Waterloo`**, Canada. Worked on XML indexing and query rewriting, with Tamer Özsu

## Awards

2021 *VLDB 2021 Distinguished Reviewer Award*

2020 *SIGMOD 2020 Reproducibility Award*: Raha: A Configuration-Free Error Detection System

2018 *Best papers of ICDE 2018*: Discovering Mis-Categorized Entities

2015 *Best papers of VLDB 2015*: Lightning Fast and Space Efficient Inequality Joins

2012 *Best papers of ICDE 2012*: Incremental Detection of Inconsistencies in Distributed Data

2010 *The Best Paper Award of VLDB 2010*: Towards Certain Fixes with Editing Rules and Master Data

2009 *Best papers of ICDE 2009*: Projective Distribution of Full-Fledged XQuery

## Research at QCRI (Dec 2011–Present)

### Data preparation theory and systems

- ***Error detection.*** Collected and categorized many real-world data errors that were provided by various organizations (`PVLDB'16`). Proposed novel data quality rules and inference algorithms (`ICDE'18`, `KDD'18`, `SIGMOD'17`, `SIGMOD'19`) to capture different types of data errors.

- ***Trusted data repairing.*** Real-world users are hesitant to see their data being automatically and heuristically repaired. My works on this thread include declarative data repairing rules (SIGMOD'14, PVLDB'20), as well as methods using master data (ICDE'15) or knowledge bases (SIGMOD'15, ICDE'17), and having human-in-the-loop (SIGMOD'16).
- ***A commodity data cleaning system.*** Nadeef (SIGMOD'13, SIGMOD'15) is the first commodity data cleaning platform that provides a unified programing interface for *declaratively* specifying what are data errors and (possibly) how to fix them, and a core that holistically handles the detection and repairing of errors. Nadeef has five filed US/EU patents, based on which we made a startup attempt.
- ***Data preparation as a service.*** Collaborating with **MIT**, we are building Data Civilizer (CIDR'17) with a suite of prebuilt tools for end-to-end data preparation. During interacting with real-world scenarios, we also developed the following components.
  - *Data discovery* in a data lake (ICDE'18), driven by scenarios from Merck and Scotiabank.
  - *Interpretable entity resolution*: used program synthesis by examples to discover entity matching (PVLDB'18) and consolidation (ICDE'19) rules, based on use cases from TAMR.
  - *Relational table storage and query co-optimization:* developed deductive program synthesis algorithms for co-optimizing data storage and query plan (OOPSLA'20).
  - *Data debugging*: studied the problem of debugging data (not code) issues (CIDR'20) in data science pipelines, by working with Massachusetts General Hospital (MGH), Intel MIT lab, and All Chicago (an organization that helps homeless people in Chicago).
- ***Collaborative data preparation.*** Collaborating with **UW-Madison**, we are building crowd-in-the-loop data preparation systems. *CoClean* [SIGMOD Demo'20] is an Overleaf-like platform, built on top of Python Pandas DataFrame, which enables multiple users to collaboratively clean the same data set.

— **Deep learning for data preparation**

- ***Relational pre-trained transformer (RPT).*** Designed RPT (PVLDB'21), a Transformer-based encoder-decoder architecture that is pre-trained for a *tuple-to-tuple* model, and can be fine-tuned for *tuple-to-X* models ("*X*" could be tuple, token, label, JSON, and so on), with the goal to support a wide range of data preparation tasks such as data cleaning, auto-completion, schema matching, entity resolution, value normalization, and data transformation.
- ***Entity resolution.*** Devised deep learning-based methods for blocking (PVLDB'21) and entity matching (PVLDB'18).

— **Data preparation for machine learning**

- ***Adaptive data augmentation:*** Employed Generative Adversarial Networks (GANs) for adapting the train data set to be "similar" to the test data for enhancing supervised machine learning, when the train data set and the test data set have different missing values patterns (PVLDB'21).

— **Data visualization for data preparation**

- ***Visualization recommendation:*** Developed DeepEye (ICDE'18, SIGMOD'18 demo) that trains a binary classifier for deciding good/bad visualizations, and a ranking model to rank good visualizations. This technique has been used by Tencent and ByteDance.
- ***Natural language to visualization translation (NL2VIS)*:** Produced the first NL2VIS benchmark (SIGMOD'21) that can be used to train machine translation models. Based on this benchmark, we are the first to apply Transformer-based machine translation model on NL2VIS and demonstrate that it significantly outperforms the state of the art (IEEE VIS'21).
- ***COVID-19 dashboards.*** Developed and led the following dashboards. (1) *Qatar situation dashboard:* used by MOI Qatar; showcased on AlJazeera, Turkish TV. (2) *COVID mobility analysis:* used by MOPH Qatar, Kuwait Health Ministry, and Nigerian National Bureau of Statistics. (3) *COVID data and mobility analysis in China* (IEEE Data Eng. Bull.'20). In early 2020, we built a COVID-19 dashboard that attracted millions of visits, and worked with China Mobile to visualize and analyze the trajectories of infected persons in Beijing.

## Research at University of Edinburgh (Feb 2010–Dec 2011)

Data Cleaning

Worked on data cleaning using master data (PVLDB'10, the Best Paper Award), interacting different types of data quality rules (SIGMOD'11), incrementally detecting errors in distributed data (ICDE'12), and inferring data currency and consistency for conflict resolution (ICDE'13).

Graph Algorithms

Worked on graph pattern matching algorithms (PVLDB'10) and then added regular expressions to graph pattern queries (ICDE'11).

### Research at CWI (Feb 2008–Jan 2010)

MonetDB

Worked on efficiently supporting updates in column-stores using packed memory arrays. building space-economical *Q*-gram index for exact string matching over a 400+GB data set (CIKM'09), and enabling efficient distribution of full-fledged XQuery on top of MonetDB/XQuery (ICDE'09), for supporting use cases in the Netherlands Forensic Institute.

## Teaching and Mentoring Experience

### Mentored Interns and Postdocs, QCRI

| | | |
|---|---|---|
| Hakim Qahtan | Ph.D., KAUST, Saudi Arabia (now assistant professor at Utrecht) | 2017/09-2020/08 |
| Jinsong Guo | Ph.D., University of Oxford, UK | 2017/03-2017/09 |
| Sibo Wang | Ph.D., NTU, Singapore (now assistant professor at CUHK) | 2016/06-2016/11 |
| Dong Deng | Ph.D., Tsinghua University, China (now assistant professor at Rutgers) | 2016/06-2016/08 |
| Sourav Medya | Ph.D., UC Santa Barbara, US (now research assistant professor at Northwestern) | 2016/06-2016/08 |
| Qing Chen | Master, Fudan University, China (now Ph.D. at Zurich University) | 2015/07-2016/04 |
| Jian He | Master, Tsinghua University, China (now at Google) | 2014/11-2015/02 |
| Matteo Interlandi | Ph.D., University of Modena, Italy (now at MSR) | 2014/03-2014/05 |
| Chu Xu | Ph.D., University of Waterloo, Canada (now assistant professor at Georgia Tech) | 2013/05-2014/07 |
| Jiannan Wang | Ph.D., Tsinghua University, China (now associate professor at Simon Fraser) | 2012/12-2013/02 |
| Yu Tang | Master, Hong Kong University, HK | 2012/11-2013/01 |
| Amr Ebaid | Ph.D., Purdue University, US (now at Google) | 2012/04-2013/01 |
| Ahmed Eldawy | Ph.D., University of Minnesota, US (now assistant professor at UC Riverside) | 2012/01-2012/05 |
| Michele Dallachiesa | Ph.D., University of Trento, Italy | 2012/01-2012/05 |

### Teaching, University of Edinburgh, UK (Tutorials)

| | |
|---|---|
| Applied Databases | 2010/09-11 |

### Teaching, The Chinese University of Hong Kong, Hong Kong (Tutorials)

| | |
|---|---|
| Digital Logical and Systems | 2006/09-12, 2007/09-12 |
| Fundamentals of Information Systems | 2004/09-12, 2006/01-05 |
| Information Systems Design & Analysis | 2005/01-05 |

## Selected Professional Activities and Services

PC Member

SIGMOD Exhibition Chair (2021), SIGMOD (2015, 2017–2020, 2022), PVLDB (2015, 2019–2021), KDD (2019–2021), CHI (2021), IEEE VIS (2021), ICDE (2013, 2018), EDBT (2017), SDM (2017), CIKM (2011, 2012)

Journal Reviewer

VLDB Journal (2009 – 2011, 2017, 2020 – 2021), TKDE (2007, 2011, 2012, 2016, 2018, 2020 – 2021), TKDD (2012), TODS (2013), TWEB (2012, 2015)

## Selected Publications, Tutorials, Patents, and Grants

### Data preparation theory and systems

[1] Abdulhakim Qahtan, **Nan Tang**, Mourad Ouzzani, Yang Cao, and Michael Stonebraker. *Pattern Functional Dependencies for Data Cleaning*. PVLDB 2020.

[2] John K. Feser, Samuel Madden, **Nan Tang**, and Armando Solar-Lezama. *Deductive Optimization of Relational Data Storage*. OOPSLA 2020.

[3] El Kindi Rezig, Lei Cao, Giovanni Simonini, Maxime Schoemans, Samuel Madden, Mourad Ouzzani, **Nan Tang**, and Michael Stonebraker. *Dagger: A Data (not code) Debugger*. CIDR 2020.

[4] Mashaal Musleh, Mourad Ouzzani, **Nan Tang**, and AnHai Doan. *CoClean: Collaborative Data Cleaning*. SIGMOD demo, 2020.

[5] Abdulhakim A. Qahtan, Ahmed Elmagarmid, Raul Castro Fernandez, Mourad Ouzzani, and **Nan Tang**. *FAHES: A Robust Disguised Missing Values Detector*. KDD, 2018.

[6] Rohit Singh, Vamsi Meduri, Ahmed Elmagarmid, Samuel Madden, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Armando Solar-Lezama, and **Nan Tang**. *Synthesizing Entity Matching Rules by Examples*. PVLDB, 2018.

[7] Shuang Hao, **Nan Tang**, Guoliang Li, and Jianhua Feng. *Discovering Mis-Categorized Entities*. ICDE, 2018.

[8] Saravanan Thirumuruganathan, Laure Berti-Equille, Mourad Ouzzani, Jorge-Arnulfo Quiane-Ruiz, and **Nan Tang**. *UGuide – User-Guided Discovery of FD-Detectable Errors*. SIGMOD, 2017.

[9] Shuang Hao, **Nan Tang**, Guoliang Li, Jian Li, and Jianhua Feng. *Cleaning Relations using Knowledge Bases*. ICDE, 2017.

[10] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibo Wang, Michael Stonebraker, Ahmed Elmagarmid, Ihab F. Ilyas, Samuel Madden, Mourad Ouzzani, and **Nan Tang**. *The Data Civilizer System*. CIDR, 2017.

[11] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and **Nan Tang**. *Interactive and Deterministic Data Cleaning: A Tossed Stone Raises a Thousand Ripples*. SIGMOD, 2016.

[12] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and **Nan Tang**. *Detecting Data Errors: Where are we and what needs to be done?* PVLDB, 2016.

[13] Zuhair Khayyat, Ihab F. Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, **Nan Tang**, and Si Yin. *BigDansing: A System for Big Data Cleansing*. SIGMOD, 2015.

[14] Xu Chu, John Morcos, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, **Nan Tang**, and Yin Ye. *KATARA: Reliable Data Cleaning with Knowledge Bases and Crowdsourcing*. SIGMOD, 2015.

[15] Matteo Interlandi, and **Nan Tang**. *Proof Positive and Negative in Data Cleaning*. ICDE, 2015.

[16] Jiannan Wang, and **Nan Tang**. *Towards Dependable Data Repairing with Fixing Rules*. SIGMOD, 2014.

[17] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and **Nan Tang**. *NADEEF: A Commodity Data Cleaning System*. SIGMOD, 2013.

[18] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Interaction Between Record Matching and Data Repairing*. SIGMOD, 2011.

[19] Wenfei Fan, Jianzhong Li, Shuai Ma, **Nan Tang**, and Wenyuan Yu. *Towards Certain Fixes with Editing Rules and Master Data*. PVLDB, 2010. (**The best paper award**)

— **Deep learning for data preparation**

[20] Saravanan Thirumuruganathan, Han Li, **Nan Tang**, Mourad Ouzzani, Yash Govind, Derek Paulsen, Glenn Fung, and AnHai Doan. *Deep Learning for Blocking in Entity Matching: A Design Space Exploration*. PVLDB, 2021.

[21] **Nan Tang**, Ju Fan, Fangyi Li, Jianhong Tu, Xiaoyong Du, Guoliang Li, Sam Madden, and Mourad Ouzzani. *RPT: Relational Pre-trained Transformer Is Almost All You Need for Democratizing Data Preparation*. PVLDB, 2021.

[22] Saravanan Thirumuruganathan, **Nan Tang**, Mourad Ouzzani, and AnHai Doan. *Data Curation with Deep Learning*. EDBT, 2020.

[23] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzzani, and **Nan Tang**. *Distributed Representations of Tuples for Entity Resolution*. PVLDB, 2018.

### Data preparation for machine learning

[24] Tongyu Liu, Yinqing Luo, Ju Fan, **Nan Tang**, Guoliang Li, and Xiaoyong Du. *Adaptive Data Augmentation for Supervised Learning over Missing Data*. PVLDB, 2021.

[25] Jianbin Liu, Fu Zhu, Chengliang Chai, Yuyu Luo, and **Nan Tang**. *Automatic Data Acquisition for Deep Learning*. VLDB demo, 2021.

### Data visualization for data preparation

[26] Yuyu Luo, **Nan Tang**, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. *Natural Language to Visualization by Neural Machine Translation*. IEEE VIS, 2021.

[27] Yuyu Luo, **Nan Tang**, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. *Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks*. SIGMOD, 2021.

[28] Yuyu Luo, **Nan Tang**, Guoliang Li, Tianyu Zhao, Wenbo Li, and Xiang Yu. *DEEPEYE: A Data Science System for Monitoring and Exploring COVID-19 Data*. IEEE Data Engineering Bulletin, 2020. (Invited)

[29] Yuyu Luo, Chengliang Chai, Xuedi Qin, **Nan Tang**, and Guoliang Li. *Interactive Cleaning for Progressive Visualization through Composite Questions*. ICDE, 2020.

[30] Yuyu Luo, Wenbo Li, Tianyu Zhao, Xiang Yu, Lixi Zhang, Guoliang Li, and **Nan Tang**. *DeepTrack: Monitoring and Exploring Spatio-Temporal Data (A Case of Tracking COVID-19)*. VLDB demo, 2020.

[31] Xuedi Qin, Yuyu Luo, **Nan Tang**, and Guoliang Li. *Making Data Visualization More Efficient and Effective: A Survey*. VLDBJ, 2020.

[32] Xuedi Qin, Yuyu Luo, **Nan Tang**, and Guoliang Li. *DeepEye: Towards Automatic Data Visualization*. ICDE, 2018.

### Tutorials

[33] **Nan Tang**, Eugene Wu, and Guoliang Li. *Towards Democratizing Relational Data Visualization*. SIGMOD tutorial, 2019.

### Patents

[34] *Dependable Data Repairing with Fixing Rules*. QCRI, HBKU (PCT/EP2013/052476).

[35] *Towards Dependable Data Repairing with Fixing Rules*. QCRI, HBKU (PCT/EP2014/052494).

[36] *KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing*. QCRI, HBKU (PCT/GB2014/051670).

[37] *NADEEF: A Holistic and Extensible Data Cleaning Platform*. QCRI, HBKU (PCT/EP2012/062446).

[38] *Generalized Data Cleaning using SAT-Solvers*. QCRI, HBKU (PCT/EP2012/062445).

### *Grants*

[39] *Credible Open Knowledge Network* (NSF grant #1937143). Start date: September 1, 2019. End Date: May 31, 2021. Prof. Chengkai Li from **University of Texas at Arlington** is the PI and I serve as a strategic partner.

[40] *Effective and Efficient Data Quality Management for Data Lakes* (Australian Research Council: DP210103593). From 2021 to present. Professor Wei Wang from **University of New South Wales** is the PI and I serve as a co-PI.

## Invited Talks

2020/05 *Data Visualization and Exploration of COVID-19 data*, QCRI lectures on the use of AI techniques for COVID-19, Qatar. (Reported by Gulf Times.)

2019/10 *Data Preparation meets Data Visualization*, at Northeastern University, US.

2016/10 *Mind Your Analytics, Clean Your Data*, at Harvard University, US.

2016/03 *Graph Stream Summarization*, at MIT, US.

2015/12 *Trusted Data Cleaning*, at KAUST, Saudi Arabia.

2014/09 *Big Data Cleaning*, at Asia-Pacific Web Conference 2014, Distinguished Lecturer series.